Vol.2018-MUS-120 No.8
2018/8/22

WaveNetを用いた楽譜情報に基づく歌唱F0軌跡の生成

和田 雄介[1,a)]　錦見 亮[1,b)]　中村 栄太[1,c)]　糸山 克寿[1,d)]　吉井 和佳[1,e)]

概要：本稿では，音符系列（楽譜）から，WaveNet と呼ばれる深層自己回帰モデルを用いて歌唱 F0 軌跡を生成する手法を示す．歌唱 F0 軌跡には，ビブラートやポルタメントなど，時間・周波数方向の複雑な変動が含まれる．従来は，このような変動を表現するのに隠れマルコフモデル（HMM）が用いられていたが，歌唱 F0 軌跡の複雑な変動を正確に捉えるためには，より表現力の高いモデルが必要である．この問題を解決するため，近年，深層自己回帰モデル WaveNet を用いて，楽譜と歌詞から歌唱 F0 軌跡を生成する手法が提案された．この手法を基に，本研究では，WaveNet の歌詞情報なしに歌唱 F0 軌跡を生成する能力を調査する．提案手法では，WaveNet による歌唱 F0 軌跡の生成を，音符系列および楽譜から抽出した特徴量によって条件付ける．また，オリジナルの WaveNet では学習時にクロスエントロピー誤差が用いられているが，生成される歌唱 F0 軌跡の自然さを高めるため，提案手法では，正解 F0 軌跡と予測との平均二乗誤差に比例する重みがついたクロスエントロピーを損失関数として用いる．実験の結果，楽譜から抽出した特徴量の追加および損失関数の変更が，どちらも生成された歌唱 F0 軌跡の品質向上に寄与することを示した．

1. はじめに

歌唱表現は，ビブラートやポルタメントなどの音高変動，音量の変化や声質から成り，歌声を特徴付ける上で重要である．特に，音高の変動は多様な歌唱表現を含み，歌唱 F0 軌跡の生成モデルは，自然かつ表現豊かな歌声の合成に有用である．このようなモデルは，歌唱スタイルの転写や，商用歌声合成ソフトウェア VOCALOID [1] に代表される歌声合成器のパラメータの自動調整に応用可能である．また，F0 軌跡生成手法を声質変換手法 [2–4] と組み合わせることで，ある歌声を，任意の別の歌手による歌唱に変換することができる．

従来の歌唱 F0 軌跡生成手法は，二次の線形システム [5] や HMM [6, 7]，ガウス過程回帰の混合エキスパートモデル [8] といった明示的なモデリングに基づいている．これらのモデルは，ある歌手に特有の音高変動を解析するのに有用であるが，自然な歌唱 F0 軌跡を生成するには，より表現力の高いモデルが必要である．近年提案された深層自己回帰モデル [9, 10] は，非線形な表現を学習でき，この問題を解決できると期待される．WaveNet [10] は，音声波形をモデル化するために提案された畳み込みニューラルネットワークであり，声質変換 [4] やテキスト音声合成 (text-to-speech; TTS) [11]，楽器音合成 [12] など，様々なタスクに応用されている．

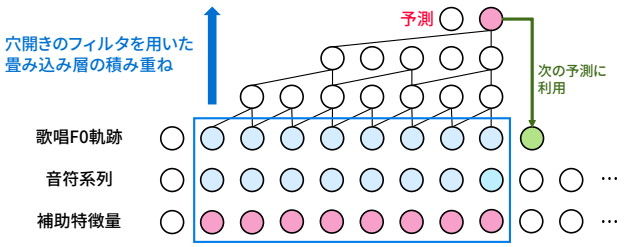

図 1: WaveNet を用いた音符系列に対する歌唱 F0 軌跡の生成の概念図．現在のフレームより前の F0 軌跡および補助特徴量が，dilated convolution 層の積み重ねに入力される．現在のフレームの歌唱 F0 の値は，自己回帰によって予測され，その値は次のフレームの予測に利用される．

歌声合成のための，WaveNet に基づく歌唱 F0 軌跡生成手法が近年提案された [13]．このモデルは，音符系列および歌詞から抽出した音素情報に基づいて，歌唱 F0 軌跡を生成する．歌唱 F0 軌跡は音素の影響を受けるため，歌唱スタイル変換への応用を考えたとき，歌手特有の音高変動を，音素情報なしに適切にモデル化できるかどうかが問題である．

本稿では，WaveNet の，歌詞情報を用いずに音符系列から歌唱 F0 軌跡を生成する能力を調査する (図 1)．歌詞情報なしに音符系列のみを扱うことで，あらゆる言語の曲に

[1] 京都大学 大学院情報学研究科
[a)] wada@sap.ist.i.kyoto-u.ac.jp
[b)] nishikimi@sap.ist.i.kyoto-u.ac.jp
[c)] enakamura@sap.ist.i.kyoto-u.ac.jp
[d)] itoyama@sap.ist.i.kyoto-u.ac.jp
[e)] yoshii@sap.ist.i.kyoto-u.ac.jp

対応できる．また，歌詞に関わらず，音型に依存して現れる歌唱表現を捉える狙いがある．TTS の手法 [14,15] を参考に，提案手法では，(I) 現在のフレームより後の音符系列，(II) 現在のフレームの音符内相対位置，(III) 現在のフレームが属する音符のフレーム単位の長さ，(IV) 歌手コードの 4 つの特徴量を楽譜情報から抽出し，これらを用いて WaveNet による生成を条件付ける．特徴量 (I) について，人間は，次に歌うべき音符を知ることで，滑らかに歌うことができるため，生成される歌唱 F0 軌跡の滑らかさを増すために導入する．特徴量 (II) および (III) は，音符の長さや音符内位置に依存する歌唱表現を捉えるために導入する．例えば，ポルタメントは音符同士の境目に現れやすく，ビブラートは比較的長い音符の終わり付近に現れやすい．特徴量 (IV) は，ある歌手に特有の歌唱表現を捉えるために導入する．

また，WaveNet の学習時に用いる損失関数の変更も行った．オリジナルの WaveNet では，損失関数としてクロスエントロピーが用いられる．クロスエントロピー関数は，あらゆる予測誤りに対して同じ損失の値を返す．WaveNet を歌唱 F0 軌跡の生成に利用する際，クロスエントロピー関数を用いると，正解 F0 の値から離れた予測を抑制できない．よって提案手法では，クロスエントロピー関数に，予測値が正解 F0 の値から離れるほど値が増加する重みを掛けた関数を損失関数として用いる．

本研究の主な貢献は，歌詞情報を使わず楽譜から抽出できる特徴量を導入したことと，オリジナルの WaveNet で用いられるクロスエントロピー関数を，歌唱 F0 軌跡の生成により適した形に変更したことである．これらの手法の有効性を，生成された歌唱 F0 軌跡と正解 F0 軌跡の二乗平均平方根 (RMSE) の計測によって評価した．その結果，提案手法における RMSE の値は，オリジナルの WaveNet より小さくなることが確認された．これより，上に示した 2 つの手法が，どちらも生成される歌唱 F0 軌跡の品質向上に寄与することが示された．

2. 関連研究

本章では，歌声および音声の合成と，F0 軌跡のモデル化に関する研究を概観する．

2.1 歌声合成

歌声合成は盛んに研究されている [1,5,6,8,13,16–19]．あらかじめ音素ごとに用意した歌声の素片を組み合わせる手法 [1,16,17] は，基本的ながら高品質な歌声を合成できる．このうち，商用歌声合成ソフトウェア VOCALOID [1] は，楽曲制作に広く用いられている．Bonada ら [16] は，1 人の歌手からなる母音と子音の歌声データベースをそれぞれ作成し，合成に用いた．Ardaillon ら [17] は，歌唱 F0 軌跡をビブラートやオーバーシュートなどに分類し，それぞれを B スプライン曲線を用いてモデル化した．歌声および歌唱 F0 軌跡合成のための統計的手法も，数多く提案されている [5–8,18]．Sinsy [18] は，歌詞・音高・音長を同時に扱う HMM を用いた統計的歌声合成手法である．また，歌唱 F0 軌跡の明示的な生成モデルとして，二次の線形システム [5] や，HMM [6,7]，ガウス過程回帰の混合エキスパートモデル [8] が提案されている．

ディープニューラルネットワーク (DNN) に基づく歌声合成も提案されている [13,19]．Nishimura ら [19] は，Sinsy [18] における HMM ベースのモデルを，全結合ニューラルネットワークに置き換え，合成された歌声の品質を向上させた．Blaauw ら [13] は，音素の発音タイミング，音程，音色の 3 つを WaveNet を用いてモデル化し，state-of-the-art な品質の歌声合成を達成した．

2.2 テキスト音声合成

TTS は，歌声合成と同様に活発に研究されている [11,14,15,20–23]．音素片を合成する手法 [20,21] は，1990 年代から研究され，高品質な音声の合成を可能にした．HMM に基づく統計的手法も提案されている [22,23]．Zen ら [22] は，HMM の拡張として，ボコーダによる音声合成のための静的・動的特徴量の相関を明示的に扱うトラジェクトリ HMM を提案した．Kameoka ら [23] は，声帯振動を表現する 2 次の線形システムである藤崎モデル [24] の確率的な定式化として，HMM に基づく音声 F0 軌跡の生成モデルを提案した．

DNN に基づく end-to-end な合成手法が，近年提案されている [11,14,15]．Fan ら [15] と Zen ら [14] は，ボコーダによる音声合成のためのパラメータを LSTM を用いて生成する手法を提案した．これらのモデルでは，入力のテキストから抽出された，音素レベルの言語的特徴 (音素 ID，強勢，単語内の音素数，音節位置) や，あるフレームの音素内位置およびその音素の長さといった特徴量を LSTM の入力とする．Shen ら [11] は，LSTM に基づく特徴量生成器と，WaveNet に基づくボコーダを組み合わせた．

3. 提案手法

本章では，まず WaveNet の定式化について説明する．その後，WaveNet に基づいて楽譜情報から歌唱 F0 軌跡を生成する提案手法について説明する．

3.1 WaveNet

本研究では，WaveNet は音符系列から歌唱 F0 軌跡を生成するのに用いられる（図 2）．WaveNet は，入力された時系列データ $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ の同時確率

$$p(\boldsymbol{x}) = \prod_{t=1}^{T} p(\boldsymbol{x}_t | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}). \quad (1)$$

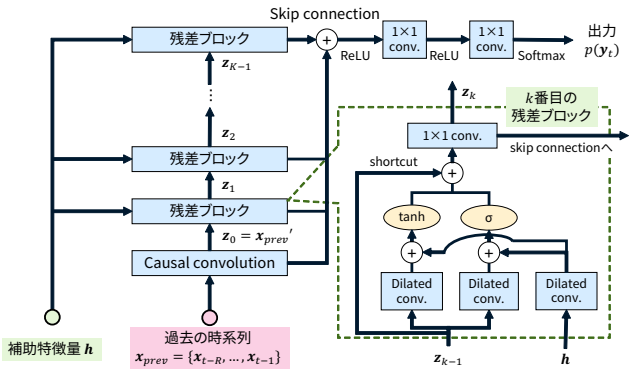

図 2: WaveNet の概要図.

を計算する．時系列データ $\boldsymbol{x}$ は，しばしば one-hot vector 形式で表現される．ネットワークの大きさは有限であり，WaveNet が実際に考慮できるサンプル数には限りがある．そのため，WaveNet は，式 (1) で表される同時確率を，

$$p(\boldsymbol{x}) \approx \prod_{t=1}^{T} p(\boldsymbol{x}_t | \boldsymbol{x}_{t-R}, \boldsymbol{x}_{t-R+1}, \cdots, \boldsymbol{x}_{t-1}) \qquad (2)$$

によって近似する．式 (2) 中の R は，WaveNet が考慮できるサンプル数であり，受容野と呼ばれる．

WaveNet が，過去の R 個のサンプル $\boldsymbol{x}_{\text{prev}} = \{\boldsymbol{x}_{t-R}, \ldots, \boldsymbol{x}_{t-1}\}$ から，現在のフレーム $\boldsymbol{x}_t$ の出力確率を計算する方法について説明する．式 (2) で表される同時確率は，残差ブロックと呼ばれる構造の積み重ねによって表現される．残差ブロックとは，3 つの 1 次元 dilated convolution (DC) 層を含み，それらの出力を 2 つの非線形な活性化関数を経て統合し出力する構造である．WaveNet に入力された $\boldsymbol{x}_{\text{prev}}$ は，1×1 の (フィルタサイズ 1 かつシフトサイズ 1 の) 1 次元 causal convolution 層を経由して $\boldsymbol{x}'_{\text{prev}}$ に変換されたのち，最初の残差ブロックに入力される．Causal convolution とは，過去の情報のみを考慮した畳み込み演算のことである．k 番目の残差ブロックの出力 $\boldsymbol{z}_k (k = 0, 1, \ldots, K)$ は，$\boldsymbol{z}_0 = \boldsymbol{x}'_{\text{prev}}$ として，

$$\boldsymbol{z}_k = \tanh(W_{f,k} * \boldsymbol{z}_{k-1}) \odot \sigma(W_{g,k} * \boldsymbol{z}_{k-1}) \qquad (3)$$

と表される．ここで，$*$ は畳み込み演算，$\odot$ は要素積，$W_{f,k}$ および $W_{g,k}$ はそれぞれ k 番目の DC 層のフィルタ，$\tanh(\cdot)$ および $\sigma(\cdot)$ はそれぞれハイパボリックタンジェント関数およびシグモイド関数を表す．全ての残差ブロックの出力は，1×1 畳み込みを経た後 skip connection によって統合され，WaveNet の最終的な出力は，softmax 関数による $\boldsymbol{x}_t$ の各要素の生起確率となる．

各残差ブロック中の DC 層の dilation（穴開き）の大きさは，

$$1, 2, 4, \ldots, 512, 1, 2, 4, \ldots, 512, 1, 2, 4, \ldots, 512.$$

のように，1 から始まって層が 1 つ進むごとに 2 倍され，

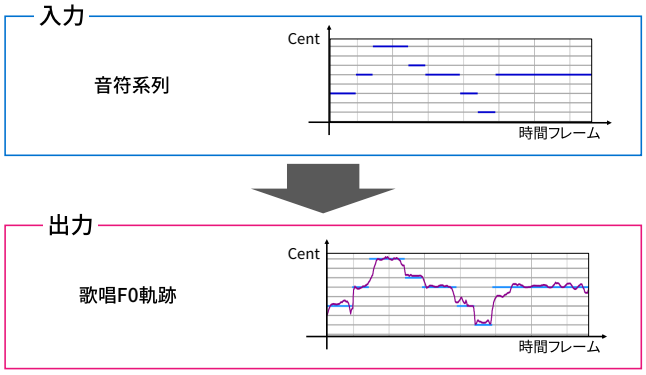

図 3: 提案手法の問題設定．入力された音符系列に対して，歌唱 F0 軌跡が出力される．

さらにある値までの dilation の系列が数回繰り返されることが多い．

残差ブロックの個数を K とし，K 番目の残差ブロック中の DC 層の dilation の大きさを d_K，dilation の系列の繰り返し回数を B としたとき，受容野 R は，

$$R = 2^{d_K} \cdot B. \qquad (4)$$

と計算される．式 (4) より，層数の増加に対して dilation の大きさを指数的に増やすことで，受容野を指数的に広げられる [25]．さらに，dilation を繰り返すことで，モデルの非線形性および表現力がさらに増加する．

WaveNet は，補助特徴量 $\boldsymbol{h} = \{h_1, h_2, \cdots, h_T\}$ を用いて，式 (1) の同時確率を，

$$p(\boldsymbol{x}|\boldsymbol{h}) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1}, \boldsymbol{h}). \qquad (5)$$

のように条件付けられる．補助特徴量は，提案手法において音符系列から抽出した特徴量に対応する．式 (5) の条件付き確率を計算するには，式 (3) を，

$$\begin{aligned} \boldsymbol{z}_k = \tanh(W_{f,k} * \boldsymbol{z}_{k-1} + W'_{f,k} * \boldsymbol{h}) \\ \odot \sigma(W_{g,k} * \boldsymbol{z}_{k-1} + W'_{g,k} * \boldsymbol{h}) \end{aligned} \qquad (6)$$

のように書き換える．ここで，$W'_{f,k}$ および $W'_{g,k}$ は，それぞれ補助特徴量を入力とする 1×1 畳み込みのフィルタを表す．

3.2 音符系列に対する歌唱 F0 軌跡の逐次予測

本節では，WaveNet を用いて音符系列から歌唱 F0 軌跡を出力する手法について述べる (図 3)．入力音符系列は，フレーム単位の対数周波数 (単位は cent) の系列 $\boldsymbol{h} = \{\boldsymbol{h}_t\}_{t=1}^{T}$ である．出力歌唱 F0 軌跡は，対数周波数 (単位は cent) の系列 $\boldsymbol{x} = \{\boldsymbol{x}_t\}_{t=1}^{T}$ である．ただし，T は系列の個数であり，$\boldsymbol{x}_t$ および $\boldsymbol{h}_t$ は one-hot vector として表現する．

歌唱 F0 軌跡の条件付き同時確率 $p(\boldsymbol{x}|\boldsymbol{h})$ は，式 (5) に従って計算される．オリジナルの WaveNet と同様に，$\boldsymbol{x}$

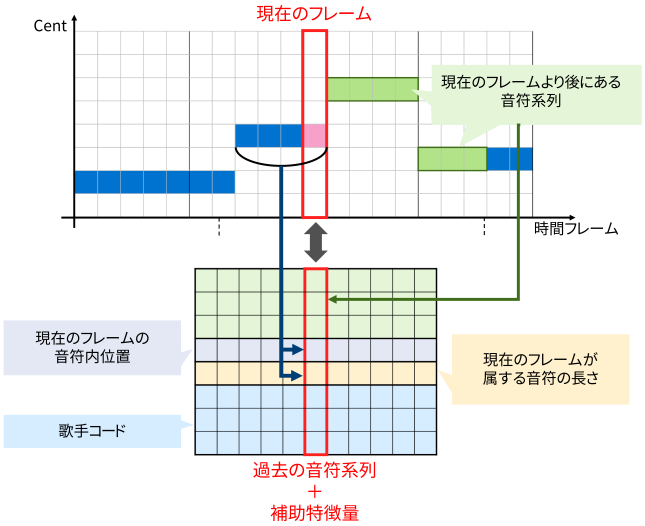

図 4: 音符系列から抽出される補助特徴量.

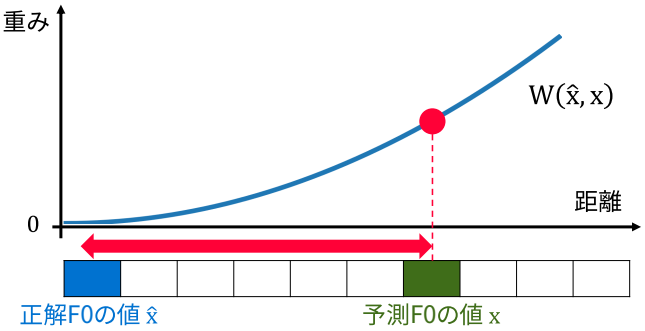

図 5: クロスエントロピーに適用する重み関数のグラフ表現.

は，1×1 畳み込みを経て最初の残差ブロックに入力される．k 番目の残差ブロックの出力は，k 番目の DC 層の出力および補助特徴量系列 $\boldsymbol{h}$ を用いて，式 (6) に従って計算される．WaveNet の最終的な出力は，softmax 関数を経た歌唱 F0 軌跡の出力確率である．式 (5) で表される同時確率の計算において，学習時には，正解 F0 の値が用いられる．これに対して，生成時には，過去に生成した F0 の値が用いられ，各時刻ごとに F0 の値が同時確率からサンプルされる．

3.3 補助特徴量

第 1 で述べたように，提案手法では，WaveNet に入力する補助特徴量として，音符系列の他に，図 4 に示した 4 つの特徴量を用いる．全ての特徴量は，1 つの系列に結合された状態で WaveNet に入力される．すなわち，式 (5) 中の $\boldsymbol{h}$ は，$\boldsymbol{c}_t$ を追加する特徴量として，$\boldsymbol{h}' = \{(\boldsymbol{h}_t, \boldsymbol{c}_t)\}_{t=1}^{T}$ に置き換えられる．図 4 に示したように，音符系列および歌手コードは one-hot vector として表され，音符内位置および音符の長さは実数値として表される．

3.4 損失関数

提案手法では，WaveNet の出力である D 次元の歌唱 F0

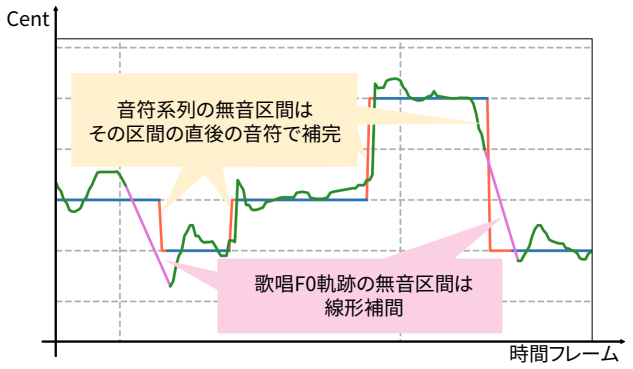

図 6: 音符系列および歌唱 F0 軌跡に含まれる，200 ミリ秒未満の無音区間の補間．歌唱 F0 軌跡に含まれる無音区間は線形補間し，音符系列に含まれる無音区間は，その区間の直後にある音符で補間した．青色の線は音符系列を，緑色の線は歌唱 F0 軌跡を表し，橙色および桃色の線はそれぞれの補間を表す．

軌跡の出力確率 $p(\boldsymbol{x})$ が，正解の F0 軌跡 $\hat{\boldsymbol{x}}$ から不自然に逸脱するのを防ぐため，以下のような重み付きクロスエントロピー関数 L を損失関数として用いる．

$$\mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = W(\hat{\boldsymbol{x}}, \boldsymbol{x}) H(\hat{\boldsymbol{x}}, \boldsymbol{x}) \tag{7}$$

ただし，

$$H(\hat{\boldsymbol{x}}, \boldsymbol{x}) = -\sum_{d=1}^{D} \hat{x}_d \log p(x_d) \tag{8}$$

は，$\hat{\boldsymbol{x}}$ と $\boldsymbol{x}$ のクロスエントロピーであり，

$$W(\hat{\boldsymbol{x}}, \boldsymbol{x}) = (d(\boldsymbol{x}, \boldsymbol{x})/100)^2 \tag{9}$$

は，$\hat{\boldsymbol{x}}$ と $\boldsymbol{x}$ の二乗誤差に比例する重み関数である．この損失関数 L は，平均二乗誤差関数として振る舞う．重み関数の係数は，損失関数の値が大きすぎると学習が失敗する現象が見られたため，経験的に決定した．歌唱 F0 軌跡の予測値は，正解の値から離れれば離れるほど不自然になると考えられる．この重み関数 $W(\hat{\boldsymbol{x}}, \boldsymbol{x})$ を用いることで，そのような逸脱を抑制できる．

4. 評価実験

本章では，提案した歌唱 F0 生成モデルの評価実験について述べる．

4.1 実験条件

RWC 研究用音楽データベース [26] のポピュラー音楽 100 曲のうち，50 曲をモデルの学習に用い，11 曲を用いて評価を行った．入力音符系列および歌唱 F0 軌跡には，アノテーションデータ [27] のうち，有音部分のみを用いた．学習時には，現在のフレームの予測に，過去の生成結果ではなくアノテーションデータを用いた．これに対して，生

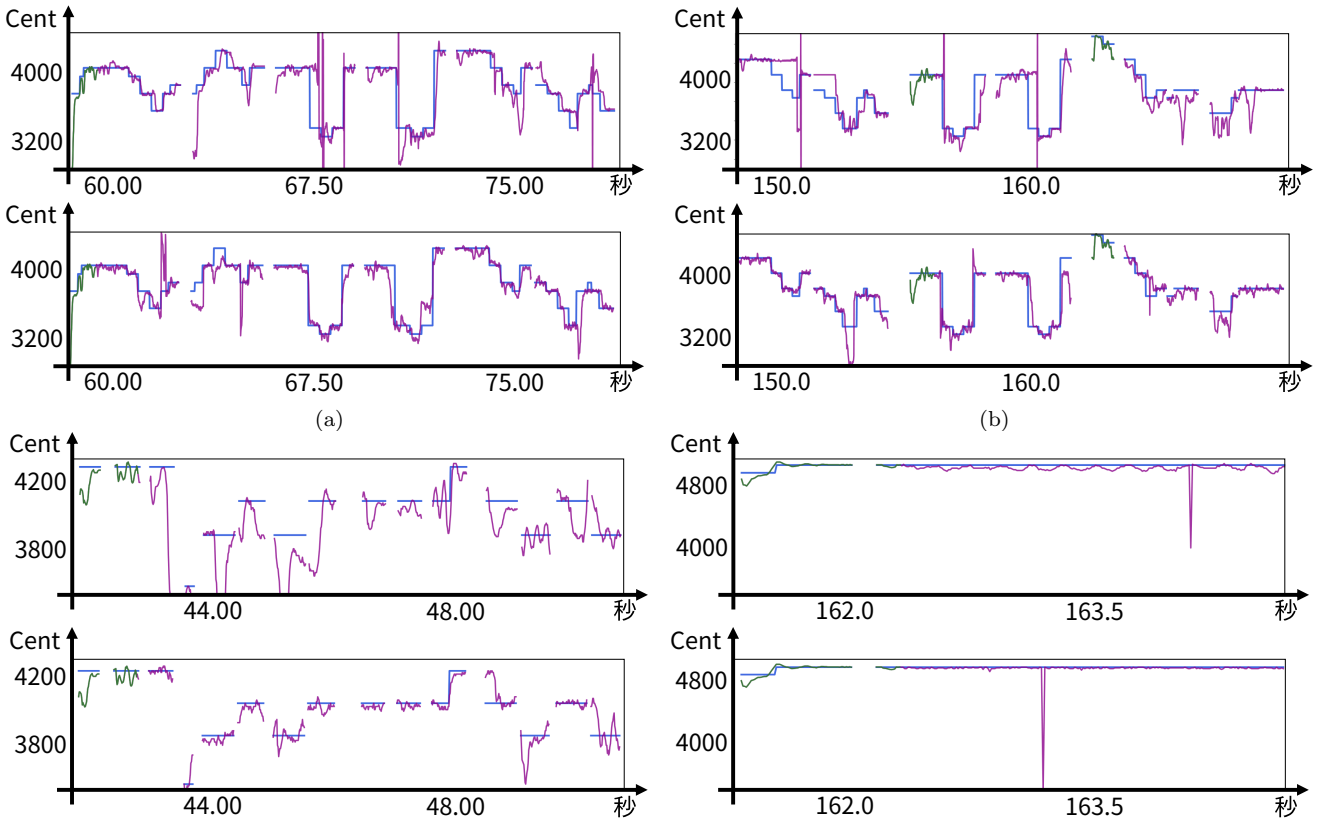

図 7: オリジナルの WaveNet および提案手法を用いて生成された歌唱 F0 軌跡の例．青色の線は入力音符系列を，紫色の線は生成された歌唱 F0 軌跡を表す．各図のうち，上側にはオリジナルの WaveNet による生成結果を示し，下側には提案手法による生成結果を示す．

成時には，過去の生成結果を用いて自己回帰予測を行った．F0 軌跡の生成の際の初期値には，全ての要素が 0 のベクトルの系列を用いた．

ピッチシフトおよび無音部分の補間によって，学習データの水増しを行った．音符系列および歌唱 F0 軌跡に含まれる，200 ミリ秒未満の無音区間を，図 6 のように補間した．また，各曲の音符系列および歌唱 F0 軌跡は，$\{-1200, -1100, \ldots, 1200\}$ の範囲からランダムに選ばれた値の分だけピッチシフトし，学習データに加えた．学習に用いられる歌唱 F0 軌跡の多様性を増すため，歌唱 F0 軌跡 $\boldsymbol{x}$ に，平均が 0，分散が半音 (100cent) のガウス分布 $\mathcal{N}(0, 100)$ に従うノイズ ϵ を加えたデータ

$$\boldsymbol{x}' = \boldsymbol{x} + \epsilon \tag{10}$$

を用意し，$\boldsymbol{x}'$ を学習に用いた．

歌唱 F0 の値は，C2 から C6 までの範囲にあるもののみを 10cent 間隔で離散化し，それ以外は無音として扱った．音符の値は，同様の範囲にあるもののみを 100cent 間隔で離散化した．このようにして離散化された歌唱 F0 軌跡および音符系列を，それぞれ 481 次元と 49 次元の one-hot vector に変換した．WaveNet に入力する補助特徴量のうち，現在のフレームより先にある音符系列については，50 サンプル分 (0.5 秒分) を用いた．学習に用いたデータセットは，74 人の歌手による歌唱が含まれている．

提案手法で用いた WaveNet は，15 層の DC 層を含み，その dilation の大きさは，入力に近い層から順に ..., 16, 1, 2, ..., 16, 1, 2, ... とした．この WaveNet の受容野は，式 (4) より 96 サンプルである．各残差ブロック内の DC 層および 1×1 畳み込み層のチャネル数は，64 とした．また，skip connection と最終的な出力の間にある 1×1 畳み込み層のチャネル数は，すべて 1024 とした．パラメータの更新は，128 サンプルを 1 ミニバッチとして，ハイパーパラメータ $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$ の Adam [28] によって行った．

提案手法によって生成された歌唱 F0 軌跡の品質を評価するため，生成された歌唱 F0 軌跡と，歌唱 F0 軌跡のアノテーションデータの間の二乗平均平方根誤差 (RMSE) を計測した．最終的な RMSE は，評価用データセット内の全曲の RMSE を平均して算出した．

4.2 実験結果

実験結果を表 1 に示した．この結果より，損失関数の変更と補助特徴量の追加が，どちらも生成された歌唱 F0 軌跡の品質向上に寄与することが分かった．RMSE の計測

表 1: 生成された歌唱 F0 軌跡と正解 F0 軌跡との RMSE.

損失関数の変更	補助特徴量の追加	RMSE [cent]
		165.4
✓		158.2
	✓	158.1
✓	✓	**150.1**

は，歌唱 F0 軌跡の品質評価に有用であるが，歌唱 F0 軌跡が歌声の品質に与える影響を調査するには，主観評価が必要であると考えられる．そのため，今後は，生成された歌唱 F0 軌跡を元に合成した音声もしくは歌声を用いて，被験者実験を行う予定である．

オリジナルの WaveNet および提案手法を用いて生成した歌唱 F0 軌跡の例を，図 7 に示す．各図において，上側に示された歌唱 F0 軌跡はオリジナルの WaveNet による生成結果であり，下側は提案手法による生成結果である．図 7a，7b および 7c に示された 3 つの例において，上側の例では歌唱 F0 軌跡が音符系列に対して不自然に逸脱しているが，下側の例ではそのような逸脱が抑制されている．さらに，これらの図のうち下側の例では，オンセット変動やプレパレーション，オーバーシュートやアンダーシュートといった歌唱表現に対応する F0 軌跡の変動が見られる．これらの変動は，補助特徴量の追加によって現れたと考えられる．以上の結果は，提案手法において用いた損失関数の変更および補助特徴量の追加が，生成される歌唱 F0 軌跡の品質向上に寄与することを示唆している．これら 3 つの例に対して，図 7d において，下側の例では，上側の例に現れているビブラートが見られない．このように，ビブラートが現れない問題は他の例でも確認されており，補助特徴量にビブラートの有無を表す変数を追加するなどの対策が必要である．

5. おわりに

本稿では，WaveNet に基づいて，音符系列から歌詞情報なしに歌唱 F0 軌跡を生成する手法について述べた．提案手法では，WaveNet への入力に音符系列から抽出した特徴量を追加し，損失関数を変更した．実験によって，これらの手法がどちらも生成される歌唱 F0 軌跡の品質向上に寄与することを確かめた．

本研究の今後の方向として，提案手法を歌唱スタイルの変換に用いるのは興味深い．今後は，提案手法と同様のアーキテクチャを，歌唱表現において F0 軌跡と同様に重要である歌唱の音量変化のモデル化にも用いる予定である．歌唱 F0 軌跡および音量の表現モデルを組み合わせることで，ある歌手の声質はそのままに，歌唱スタイルのみを別の歌手のものに変更できると考えられる．そのようなモデルを構築するには，ある歌手に特有の歌唱スタイルを学習する必要があるが，その際データの不足が予想される．この問題の解消には，転移学習が有用であると考えられる．

謝辞 本研究の一部は，JST ACCEL No. JPMJAC1602，JSPS 科研費 No. 26700020，No. 16H01744 および No. 16J05486 の支援を受けた．

参考文献

[1] Kenmochi, H. and Ohshita, H.: VOCALOID-Commercial Singing Synthesizer Based on Sample Concatenation, *Proc. Interspeech*, pp. 4009–4010 (2007).
[2] Hsu, C., Hwang, H., Wu, Y., Tsao, Y. and Wang, H.: Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks, *Proc. Interspeech*, pp. 3364–3368 (2017).
[3] Kinnunen, T., Juvela, L., Alku, P. and Yamagishi, J.: Non-parallel Voice Conversion Using I-vector PLDA: Towards Unifying Speaker Verification and Transformation, *Proc. ICASSP*, pp. 5535–5539 (2017).
[4] Kobayashi, K., Hayashi, T., Tamamori, A. and Toda, T.: Statistical Voice Conversion with WaveNet-based Waveform Generation, *Proc. Interspeech*, pp. 1138–1142 (2017).
[5] Saitou, T., Unoki, M. and Akagi, M.: Development of an F0 Control Model Based on F0 Dynamic Characteristics for Singing-voice Synthesis, Vol. 46, No. 3, pp. 405–417 (2005).
[6] Lee, S. W., Ang, S., Dong, M. and Li, H.: Generalized F0 Modelling with Absolute and Relative Pitch Features for Singing Voice Synthesis, *Proc. ICASSP*, pp. 429–432 (2012).
[7] Ohishi, Y., Kameoka, H., Mochihashi, D. and Kashino, K.: A Stochastic Model of Singing Voice F0 Contours for Characterizing Expressive Dynamic Components, *Proc. Interspeech*, pp. 474–477 (2012).
[8] Ohishi, Y., Mochihashi, D., Kameoka, H. and Kashino, K.: Mixture of Gaussian Process Experts for Predicting Sung Melodic Contour with Expressive Dynamic Fluctuations, *Proc. ICASSP*, pp. 3714–3718 (2014).
[9] Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A. and Bengio, Y.: SampleRNN: An Unconditional End-to-End Neural Audio Generation Model, *Proc. ICLR*, pp. 1–11 (2017).
[10] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, *arXiv preprint arXiv:1609.03499*, pp. 1–15 (2016).
[11] Shen, J., Pang, R., Weiss, R., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R., Agiomyrgiannakis, Y. and Wu, Y.: Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, *Proc. ICASSP*, pp. 1–5.
[12] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K. and Norouzi, M.: Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders, *Proc. ICML*, pp. 1068–1077 (2017).
[13] Blaauw, M. and Bonada, J.: A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs, Vol. 7, No. 12, pp. 1313–1333 (2017).
[14] Zen, H. and Sak, H.: Unidirectional Long Short-term Memory Recurrent Neural Network with Recurrent Output Layer for Low-latency Speech Synthesis, *Proc. ICASSP*, pp. 4470–4474 (2015).

[15] Fan, Y., Qian, Y., Xie, F. and Soong, F.: TTS Synthesis with Bidirectional LSTM Based Recurrent Neural Networks, *Proc. Annual Conference of the International Speech Communication Association, Interspeech*, pp. 1964–1968 (2014).
[16] Bonada, J., Umbert, M. and Blaauw, M.: Expressive Singing Synthesis Based on Unit Selection for the Singing Synthesis Challenge 2016, *Proc. Interspeech*, pp. 1230–1234 (2016).
[17] Ardaillon, L., Degottex, G. and Roebel, A.: A Multi-layer F0 Model for Singing Voice Synthesis Using A B-spline Representation with Intuitive Controls, *Proc. Interspeech*, pp. 3375–3379 (2015).
[18] Saino, K., Zen, H., Nankaku, Y., Lee, A. and Tokuda, K.: An HMM-based Singing Voice Synthesis System, *Proc. Interspeech*, pp. 2274–2277 (2006).
[19] Nishimura, M., Hashimoto, K., Oura, K., Nankaku, Y. and Tokuda, K.: Singing Voice Synthesis Based on Deep Neural Networks, *Proc. Interspeech*, pp. 2478–2482 (2016).
[20] Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y. and Syrdal, A.: The AT&T Next - Gen TTS System, *Proc. Joint ASA/EAA/DAEA Meeting*, pp. 15–19 (1999).
[21] Coorman, G., Fackrell, J., Rutten, P. and Coile, B.: Segment Selection in the L&H Realspeak Laboratory TTS System, *Proc. Spoken Language Processing*, pp. 395–398 (2000).
[22] Zen, H., Tokuda, K. and Kitamura, T.: Reformulating the HMM as a Trajectory Model by Imposing Explicit Relationships between Static and Dynamic Feature Vector Sequences, Vol. 21, No. 1, pp. 153–173 (2006).
[23] Kameoka, H., Yoshizato, K., Ishihara, T., Kadowaki, K., Ohishi, Y. and Kashino, K.: Generative Modeling of Voice Fundamental Frequency Contours, Vol. 23, No. 6, pp. 1042–1053 (2015).
[24] Fujisaki, H.: A Note on the Physiological and Physical Basis for the Phrase and Accent Components in the Voice Fundamental Frequency Contour, pp. 347–355 (1988).
[25] Yu, F. and Koltun, V.: Multi-scale Context Aggregation by Dilated Convolutions, *Proc. ICLR*, pp. 1–13 (2016).
[26] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, *Proc. ISMIR*, pp. 229–230 (2003).
[27] Goto, M.: AIST Annotation for RWC Music Database, *Proc. ISMIR*, pp. 359–360 (2006).
[28] Kingma, D. and Ba, J.: Adam: A Method for Stochastic Optimization, *Proc. ICLR*, pp. 1–15 (2015).